Theoretical Computer Science 456 (2012) 89–99

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Generic subset ranking using binary classifiers☆

Zhengya Sun [a,*], Wei Jin [b], Jue Wang [a]

[a] Institute of Automation, Chinese Academy of Sciences, China

[b] Department of Computer Science, North Dakota State University, USA

ARTICLE INFO

Article history:
Received 22 April 2011
Received in revised form 6 May 2012
Accepted 19 May 2012
Communicated by R. Gavalda

Keywords:
Subset ranking
Position-sensitive measures
Regret bound

ABSTRACT

A widespread idea to attack the ranking problem is by reducing it into a set of binary preferences and applying well studied classification methods. In particular, we consider this reduction for generic subset ranking, which is based on minimization of position-sensitive loss functions. The basic question addressed in this paper relates to whether an accurate classifier would transfer directly into a good ranker. We propose a consistent reduction framework guaranteeing that the minimal regret of zero for subset ranking is achievable by learning binary preferences assigned with importance weights. This fact allows us to further develop a novel upper bound on the subset ranking regret in terms of binary regrets. We show that their ratio can be at most 2 times the maximal deviation of discounts between adjacent positions. We also present a refined version of this bound when only the quality over the top rank positions is of concern. These bounds provide theoretical support on the use of the resulting binary classifiers for solving the subset ranking problem.

1. Introduction

Supervised rank learning tasks often boil down to the problem of ordering a finite subset of instances in an observable feature space. This task is referred to as subset rank learning [12]. One straightforward and widely known solution for subset ranking has been based on a reduction to binary classification tasks considering all pairwise preferences on the subset. Numerous ranking algorithms fall within the scope of this approach, i.e., building ranking models by running classification algorithms for binary preference problems [8,9,11,13,14,16,25]. Ranking models are often evaluated by position-sensitive performance measures [20], which assign each rank position a discount factor to emphasize the quality near the top. This presumable difference poses a question on whether an accurate classifier would transfer directly into a good ranker. Applications of the aforementioned algorithms seem to support this claim. In this paper, we attempt to provide theoretical support for this phenomenon based on well-established regret transform principles [6,19], a mainstay of reduction analysis. Roughly speaking, *regret* here describes the gap between the incurred loss and the minimal loss.

Recent theoretical developments in rank learning can be tracked in two directions, one towards generalization properties [10,17], and the other towards reduction analysis which is the focus of our concern. Relevant work on reduction analysis has shown that the ranking problem can be solved robustly and efficiently with binary classification techniques. The proved regret bounds for those reductions, however, mostly focus on measures that are not position-sensitive. For example, Balcan et al. [3] proved that the regret of ranking, as measured by the Area Under the ROC Curve (AUC), is at most twice as much as that of the induced binary classification. Ailon and Mohri [1] described a randomized reduction which guarantees that the pairwise misranking regret is not more than the binary classification regret. These inspiring results lead us to seek

☆ A preliminary version of this work appeared in the Proceedings of the 24th Canadian Conference on Artificial Intelligence (CAI 2011), vol. 6657, pp. 396–407.

* Corresponding author. Tel.: +86 10 82616599.
E-mail address: zhengya.sun@ia.ac.cn (Z. Sun).

0304-3975/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2012.05.026

regret guarantees for ranking under position-sensitive measures, which have gained enormous popularity in practice, such as Average Precision (AP) [2,23], Normalized Discounted Cumulative Gain (NDCG) [15], and so on.

In [24], the authors proposed a NDCG reduction framework, and bounded the NDCG regret in terms of the importance weighted classification regret. However, the presented theoretical analysis is insufficient in backing up their arguments. Two closely related aspects for successful reduction from position-sensitive ranking to binary classification have been ignored. These are (a) guarantees that the reduction is consistent in the sense that given optimal (zero-regret) binary classifiers, the reduction can yield an optimal ranker, such that the expected position-sensitive performance measure is maximized, and (b) the regret bounds which demonstrate the decrease of classification regret may provide a reasonable approximation for the decrease of ranking regret of interest.

Our current study aims at addressing these problems. Although the first aspect has been analogously pointed out in [12], there has been no comprehensive theoretical analysis to our knowledge. We characterize each position-sensitive ranking measure by a combination of 'relevance gain' and 'position discount', and prove that under suitable assumptions, the sufficient condition on consistent reduction is given by learning pairwise preferences assigned with importance weights according to relevance gains. In particular, we derive an importance weighted loss function for the reduced binary problems that exhibits good properties in preserving an optimal ranking. Such properties provide reassurance that optimizing the resulting binary loss in expectation does not hinder the search for a zero-regret ranker, and allow such a search to proceed within the scope of off-the-shelf classification algorithms.

Subsequently, we quantify the reduction with consistency guarantee on at most how much the classification regret can be transferred into the position-sensitive ranking regret. Our regret analysis is based on the rank-adjacent transposition strategies which were first used to convert the ranking regret to multiple pairwise regrets. This, coupled with the majorization inequality proved by Balcan et al. (2007), allows us to yield an upper bound in terms of the sum of the importance weighted classification regrets over the induced binary problems. The bound is then scaled by a position-discount factor, i.e., 2 times the maximum deviation between adjacent position discounts (<1). This constant does not depend on how many instances are ranked, and can therefore be regarded as an improvement over that in the case of subset ranking using the regression approach [12]. The refined version of the regret bound is also presented when only the quality over the top rank positions is of concern. Our results reveal the underlying connection between position-sensitive ranking and binary classification, that is, the improvement of the classification accuracy can reasonably enhance the position-sensitive ranking performance.

The remainder of this paper is organized as follows. Section 2 formulates the subset ranking problem and analyzes its optimal behavior. Section 3 presents pairwise classification formulations and describes a generic algorithm for subset ranking with the binary classifier. Section 4 is devoted to the proof of our main results, and Section 5 concludes this paper.

2. The subset ranking problem

We consider the subset ranking problem described as follows. Provided with labeled subsets, the ranker learns to predict a mapping from a finite subset to an ordering over the instances in it. Each labeled subset is assumed to be generated in the form of $S = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$, where x_i is an instance in feature space $\mathcal{X}$, and the associated relevance label y_i belongs to the set $\mathcal{Y} = \{0, \ldots, l-1\}$ with $l-1$ representing the highest relevance and 0 the lowest.

2.1. Notation

We denote the finite subset as $X = \{x_i\}_{i=1}^n \in \mathcal{U}$ where $\mathcal{U}$ is the set of all finite subsets of $\mathcal{X}$, and the associated relevance label set as $Y = \{y_i\}_{i=1}^n$. For simplicity, the size of the subset n remains fixed throughout our analysis. We represent the ordering as a permutation π on $[n] = \{1, \ldots, n\}$, using $\pi(i)$ to denote the ranked position given to the instance x_i, and $\pi^{-1}(j)$ to denote the index of the instance ranked at the jth position. The set of all possible permutations is denoted as Ω. We also define an instance assignment vector $\mathbf{x} = [x_{\pi^{-1}(1)}, \ldots, x_{\pi^{-1}(n)}]$ and a relevance assignment vector $\mathbf{y} = [y_{\pi^{-1}(1)}, \ldots, y_{\pi^{-1}(n)}]$ according to π, where $\mathbf{x}_i = x_{\pi^{-1}(i)}$ represents the instance ranked at the ith position, and $\mathbf{y}_i = y_{\pi^{-1}(i)}$ represents the relevance label assigned to the instance ranked at the ith position. Based on the perfect ordering $\bar{\pi}$ which is in a non-increasing order of the relevance labels, we evaluate the quality of the estimated ordering π with position-sensitive performance measures $M_{\bar{\pi}}(\pi, Y)$, and use $M_{\bar{\pi}}@k(\pi, Y)$ to denote the quality of the top k portion of the ordering. Hereafter we will often omit the subscript $\bar{\pi}$ when we consider it fixed once and for all.

2.2. Position-sensitive performance measures

Many practical ranking measures are characterized by sensitivity to rank positions [4,7,15,26]. Unlike other ranking measures such as AUC, these measures discriminate each position i by a discount factor d_i, where $i = 1, 2, \ldots$, and allow the evaluation to concentrate on the top rank positions [23]. Some common position-sensitive ranking measures are shown below.

- Average Precision (AP) [23]. This measure is for binary relevance assessments with $\mathcal{Y} = \{0, 1\}$. $M@n(\pi, Y) = \frac{1}{\text{rel}@n(\bar{\pi}, Y)} \cdot \sum_{i=1}^n \frac{\mathbf{y}_i}{i} \cdot \text{rel}@i(\pi, Y)$, where $\text{rel}@i(\pi, Y) = \sum_{j=1}^i \mathbf{y}_j$; $d_i = \frac{1}{i}$.
- Normalized Discounted Cumulative Gain (NDCG) [15]. This measure is for multiple label relevance assessments with $\mathcal{Y} = \{0, 1, \ldots, l-1\}$, $l \geq 2$. $M@n(\pi, Y) = \frac{\text{DCG}@n(\pi, Y)}{\text{DCG}@n(\bar{\pi}, Y)}$, where $\text{DCG}@n(\pi, Y) = \sum_{i=1}^n \frac{2^{\mathbf{y}_i} - 1}{\log(1+i)}$; $d_i = \frac{1}{\log(1+i)}$.

Table 1
Four widely used position-sensitive ranking measures along with corresponding choices of relevance gain functions.

Measure	AP	NDCG	ERU	RBP
$g(y, Y)$	y	$\frac{2^y-1}{\text{DCG@}n(\bar{\pi},Y)}$	$\frac{\max(y-\gamma,0)}{\text{ERU@}n(\bar{\pi},Y)}$	y

- Expected Rank Utility (ERU) [7]. This measure is for multiple label relevance assessments with $\mathcal{Y} = \{0, 1, \ldots, l-1\}$, $l \geq 2$. $M@n(\pi, Y) = \frac{\text{ERU@}n(\pi,Y)}{\text{ERU@}n(\bar{\pi},Y)}$, where $\text{ERU@}n(\pi, Y) = \sum_{i=1}^{n} \max(\mathbf{y}_i - \gamma, 0) \cdot 2^{\frac{1-i}{\alpha-1}}$; $d_i = 2^{\frac{1-i}{\alpha-1}}$. The parameter γ is a neutral vote and parameter $\alpha > 1$ is the viewing half-life.
- Rank-Biased Precision (RBP) [21]. This measure is for binary relevance assessments with $\mathcal{Y} = \{0, 1\}$. $M@n(\pi, Y) = \sum_{i=1}^{n}(1-\beta) \cdot \beta^{i-1} \cdot \mathbf{y}_i$; $d_i = (1-\beta) \cdot \beta^{i-1}$. The parameter β is the probability with which the user progresses from one instance to the next.

The discount factors defined in the above measures are all positive and strictly decreasing, i.e., $\forall i \in [n-1]$, $d_i > d_{i+1} > 0$. When only the top k ($k < n$) instances need to be ranked correctly, $M@k$ is calculated, and d_i is set to be zero for $i > k$.

Definition 1 (*Relevance Decomposable*)**.** Given a permutation π and the label set Y, for each $y \in Y$, if a real-valued function ψ can be decomposed as $\psi(\pi, y, Y) = \varphi(\pi, Y) \cdot g(y, Y)$, where g is a non-constant function and φ is non-negative, then we say ψ is relevance decomposable.

Definition 2 (*Relevance Gain Function*)**.** Fix a position-sensitive performance measure M. Suppose there is a function ψ such that $M@n(\pi, Y) = \sum_{i=1}^{n} \psi(\pi, y_i, Y)$. If ψ is relevance decomposable, that is, $\psi(\pi, y_i, Y) = \varphi(\pi, Y) \cdot g(y_i, Y)$, then g is called a relevance gain function for M.

According to the above definitions, we then derive the relevance gain function tailored to each ranking measure, as shown in Table 1.

These measures are of general interest, and taken to be the optimization targets in the ranking problems of our concern.

2.3. Ranking formulations

In the standard supervised learning setup, the ranking problem that we are investigating can be defined as follows.

Definition 3 (*Position-Sensitive Subset Ranking*)**.** Assume that each labeled subset $S = \{(x_i, y_i)\}_{i=1}^{n} \subseteq \mathcal{X} \times \mathcal{Y}$ is generated independently at random according to some (unknown) underlying distribution $\mathcal{D}$. The ranker works with a ranking function space $\mathcal{H} = \{h : \mathcal{U} \to \Omega\}$ which maps a set $X \in \mathcal{U}$ to a permutation π, namely, $h(X) = \pi$. The position-sensitive ranking loss of the predictor h on a labeled subset $S = (X, Y)$ is defined as

$$l_{\text{rank}}(h, S) = M(\bar{\pi}, Y) - M(h(X), Y).$$

Note also there is an implicit dependency on $\bar{\pi}$ here. The learning goal is to find a predictor h so that the expected position-sensitive ranking loss with respect to $\mathcal{D}$, given by

$$\mathcal{L}_{\text{rank}}(h, \mathcal{D}) = E_{S\sim\mathcal{D}} l_{\text{rank}}(h, S), \tag{1}$$

is as small as possible.

The loss l_{rank} quantifies our intuitive notion of 'how far the predicted permutation is from the perfect permutation' based on a specific position-sensitive measure function M. The loss becomes minimum of zero when the subset X is ranked in a non-increasing order of the relevance labels in Y; maximal when in a non-decreasing order.

2.4. Optimal subset ranking

Let us rewrite (1) as

$$\mathcal{L}_{\text{rank}}(h, \mathcal{D}) = E_X L_{\text{rank}}(h, X). \tag{2}$$

where

$$L_{\text{rank}}(h, X) = E_{Y|X} l_{\text{rank}}(h, S). \tag{3}$$

To characterize the optimal ranking rule with the minimum loss in (2), it is convenient to analyze the conditional formulation (3) as a starting point.

Theorem 1. *Fix one of the measures chosen from* AP, NDCG, ERU *and* RBP, *and let $g(y, Y)$ be a relevance gain function for it (as listed in Table 1). Given a set $X \in \mathcal{U}$, we define the optimal subset ranking function $\hat{h}$ as a minimizer of the conditional expectation in (3). Let $\hat{\pi} = \hat{h}(X)$ be the permutation output by $\hat{h}$, then for any $d_{\hat{\pi}(i)} > d_{\hat{\pi}(j)}$, $i, j \in [n]$, it holds that*

$$E_{(y_i,y_j,Y)|(x_i,x_j,X)}(g(y_i, Y) - g(y_j, Y)) \geq 0.$$

Proof. Consider $h \in \mathcal{H}$, and assume that $h(X) = \pi$ where $\pi(k') = \hat{\pi}(k')$ when $k' \neq i, j$, and $\pi(i) = \hat{\pi}(j)$ and $\pi(j) = \hat{\pi}(i)$. By the definition of $\hat{h}$, we have

$$L_{\text{rank}}(h, X) - L_{\text{rank}}(\hat{h}, X) = E_{Y|X}(M(\hat{h}(X), Y)) - M(h(X), Y)) \geq 0.$$

We start with the case of NDCG, ERU, and RBP measures, and get that

$$(d_{\hat{\pi}(i)} - d_{\hat{\pi}(j)}) \cdot E_{(y_i, y_j, Y)|(x_i, x_j, X)}(g(y_i, Y) - g(y_j, Y)) \geq 0, \tag{4}$$

which implies the desired result.

Next, we consider the case of AP measure. Consider any $k \in \{1, \ldots, n-1\}$, and let $\pi(i) = k, \pi(j) = k+1$. Let $\text{rel@0}(\hat{\pi}, Y) = 0$. One can easily check that

$$(d_k - d_{k+1}) \cdot E_{(y_i, y_j, Y)|(x_i, x_j, X)}\left[(g(y_i, Y) - g(y_j, Y)) \cdot \left(\frac{\text{rel@}(k-1)(\hat{\pi}, Y) + 1}{\text{rel@}n(\bar{\pi}, Y)}\right)\right] \geq 0, \tag{5}$$

which implies that if $d_k > d_{k+1}$, then $E_{(y_i, y_j, Y)|(x_i, x_j, X)}(g(y_i, Y) - g(y_j, Y)) \geq 0$ holds. Applying this inequality recursively, the desired result follows. □

Note that the conditional probability of $g(y, Y)$ is dependent on X. Theorem 1 explicitly states that given X, the optimal subset ranking is in a non-increasing order of the conditional expectation $E_{(y_i, Y)|(x_i, X)} g(y_i, Y)$, where $i = 1, \ldots, n$. Generally speaking, finding the optimal subset ranking based on the conditional results is NP-hard due to possibly conflicting preferences in overlapping subsets [12]. We circumvent this problem by introducing some assumptions on the relevance gain functions, ensuring the equivalence of the optimal subset ranking and the optimal ranking on the universal set. Let $\phi : \mathbb{R} \times \mathbb{R} \times \mathcal{U} \to \mathbb{R}$ be a real-valued function which satisfies that $\phi(\mu_1, \mu_2, X) = 0$ if $\mu_1 - \mu_2 = 0$, and $\phi(\mu_1, \mu_2, X) \cdot (\mu_1 - \mu_2) > 0$ otherwise. We derive the following proposition inspired in part by [12].

Proposition 1. *Fix one of the measures chosen from* AP, NDCG, ERU, *and* RBP, *and let $g(y, Y)$ be a relevance gain function for it (as listed in Table 1). For any two instances $x_i, x_j \in X$, let $\triangle(x_i, x_j, X) = E_{(y_i, y_j, Y)|(x_i, x_j, X)}(g(y_i, Y) - g(y_j, Y))$. Assume that for each instance $x_i \in X$, there exists a random variable y'_i such that $\triangle(x_i, x_j, X) = \phi(E_{y'_i|(x_i, X)} y'_i, E_{y'_j|(x_j, X)} y'_j, X)$, and assume further that $Y' = \{y'_i\}_{i=1}^n$ is a set of random variables that satisfy*

$$P(Y'|X) = E_\xi \prod_{i=1}^{n} P(y'_i|x_i, \xi),$$

where ξ is a hidden random variable independent of X. Then $E_{y'_i|(x_i, X)} y'_i = E_{y'_i|x_i} y'_i$, and hence we have

$$E_X \min_h L_{\text{rank}}(h, X) = \min_h \mathcal{L}_{\text{rank}}(h, \mathcal{D}).$$

The proposition immediately follows from Theorem 1 and from the definition of ϕ. This result shows that the sufficient condition to remove set-dependency can be satisfied by using an appropriately defined feature function. One direct example to clarify this point is to assume that for each $x_i \in X$, we have $E_{(y_i, Y)|(x_i, X)} g(y_i, Y) = a(X)(E_{y'_i|x_i} y'_i)^k + b(X)$, where $k \in \mathbb{N}$, and $a(X) > 0$ and $b(X)$ are normalization/shifting factors that may depend on X. Since $\text{sign}(\alpha^k - \beta^k) = \text{sign}(\alpha - \beta)$ for $\alpha, \beta \geq 0$, $\triangle(x_i, x_j, X) = a(X)((E_{y'_i|x_i} y'_i)^k - (E_{y'_j|x_j} y'_j)^k)$ immediately implies that the optimality of the subset ranking conforms with that of the set-independent case.

3. Reductions to binary classification

In this section, we turn to the reduction method which decomposes subset ranking problems into importance weighted binary classification problems considering all weighted pairwise preferences between any two instances.

3.1. Classification formulations

In importance weighted binary classification, each instance–label pair is supplied with a non-negative weight which specifies the importance of predicting the category of this instance correctly [5]. The corresponding formulation [5,19] can be naturally extended to learn pairwise preferences, which is defined as follows.

Definition 4 (*Importance Weighted Binary Classification for Pairwise Preferences*)**.** Assume that each triple $t_{ij} = ((x_i, x_j), I(y_i > y_j), w_{ij}) \in (\mathcal{X} \times \mathcal{X}) \times \{0, 1\} \times [0, +\infty)$ is generated at random according to some (unknown) underlying distribution $\mathcal{P}$, where $I(\cdot)$ is 1 when the argument is true and 0 otherwise, and w_{ij} indicates the importance of the correct classification. The classifier works with a preference function space $\mathcal{C} = \{\sigma : \mathcal{X} \times \mathcal{X} \to \{0, 1\}\}$ which maps an ordered pair $(x_i, x_j) \in \mathcal{X} \times \mathcal{X}$ to a binary relation. The importance weighted classification loss of the predictor σ on a triple t_{ij} is defined as

$$l_{\text{class}}(\sigma, t_{ij}) = \frac{1}{2} w_{ij} \cdot I(y_i > y_j) \cdot (1 - \sigma(x_i, x_j) + \sigma(x_j, x_i)). \tag{6}$$

The learning goal is to find a predictor c such that the expected importance weighted classification loss with respect to $\mathcal{P}$, given by

$$\mathcal{L}_{\text{class}}(\sigma, \mathcal{P}) = E_{t_{ij}\sim\mathcal{P}} l_{\text{class}}(\sigma, t_{ij}), \tag{7}$$

is as small as possible.

When learning pairwise preferences, the binary classifier c decides for each ordered pair (x_i, x_j) whether x_i or x_j is preferred. In this regard, $\sigma(x_i, x_j) = 1$ means that x_i is strictly preferred to or equal to x_j, and $\sigma(x_i, x_j) = 0$ indicates the opposite preference, i.e., x_j is strictly preferred to or equal to x_i. A perfect prediction preserves the target preference between two alternatives, i.e., $y_i > y_j \Leftrightarrow \sigma(x_i, x_j) - \sigma(x_j, x_i) = 1$, and a non-zero loss is incurred otherwise. When $w_{ij} = 1$, the expected loss $\mathcal{L}_{\text{class}}$ is simply the probability that discordant pairs occur assuming that ties are broken at random.

Some early loss functions related to classification for pairwise preferences [1,3,24] are useful within a restricted hypothesis space satisfying the condition that $\forall x_i, x_j \in \mathcal{X}, \sigma(x_i, x_j) + \sigma(x_j, x_i) = 1$, that is, $x_i \neq x_j$. In order to have $\sigma(x_i, x_j) = \sigma(x_j, x_i)$, the output of σ cannot be 0/1 values any more. By Definition 4, σ acts as a 0/1 classifier for modeling pairwise preferences. If this is not the case, then the preference interpretation in the paragraph above is no longer valid. As an extension, the loss function defined in Eq. (6) relaxes these restrictions and is amenable to more general scenarios.

3.2. Ranking a subset with binary classifiers

We introduce a general framework for ranking a subset with a binary classifier, which unifies a large family of pairwise ranking algorithms such as Ranking SVMs [14], RankBoost [13] and RankNet[8]. This framework is composed of two procedures as described below.

The training procedure (Binary_Train) takes a set S of labeled instances in $\mathcal{X} \times \{0, \ldots, l-1\}$ and transforms every pair of labeled instances into two binary classification examples, each of which is augmented with a non-negative weight. By running a binary learning algorithm $\mathcal{A}$ on the transformed example set T, a classifier in the form of $c(x_i, x_j, X) : \mathcal{X}\times\mathcal{X}\times\mathcal{U} \to \{0, 1\}$ is obtained, where $x_i, x_j \in X$. The space of all possible functions c is denoted as $\tilde{\mathcal{C}}$.

Procedure 1 Binary_Train (A labeled set S of size n, a binary classification learning algorithm $\mathcal{A}$)

Set $T = \emptyset$.
for all ordered pairs(i, j) with $i, j \in [n]$, $i \neq j$:
  Set $\tilde{w}_{ij}(X, Y) = |g(y_i, Y) - g(y_j, Y)|$.
  Add to T an importance weighted example
  $((x_i, x_j, X), I(y_i > y_j), \tilde{w}_{ij}(X, Y))$.
end for
Return $c = \mathcal{A}(T)$.

Procedure 2 Rank_Predict (An instance set X, a binary classifier c)

for each $x_i \in X$:
  $f(x_i, X) = \frac{1}{2}\sum_{j\neq i}(c(x_i, x_j, X) - c(x_j, x_i, X) + 1)$, where $x_j \in X$.
end for
Sort X in a non-increasing order of $f(x_i, X)$.

We then define the induced distribution $\tilde{\mathcal{D}}$ on the binary classifier c. To generate a sample from this distribution, we first draw a randomly labeled set S from the original distribution $\mathcal{D}$, and subsequently draw uniformly from S an ordered pair (i, j) which is translated into $\tilde{t}_{ij} = ((x_i, x_j, X), I(y_i > y_j), \tilde{w}_{ij}(X, Y))$. We then determine the importance weight function $\tilde{w}_{ij}(X, Y)$.

Given two permutations π_1 and π_2, assume that for $i, j \in [n]$, $j = i+1, \pi_1(i) = \pi_2(j), \pi_1(j) = \pi_2(i)$, and $\forall k \in [n]$, $k \neq i, j$, $\pi_1(k) = \pi_2(k)$. Then if $y_i > y_j$, we have

$$\begin{aligned} M@n(\pi_1, Y) - M@n(\pi_2, Y) &= \sum_{k=1}^{n} \psi(\pi_1(k), Y)\cdot g(y_k, Y) - \sum_{k=1}^{n} \psi(\pi_2(k), Y)\cdot g(y_k, Y) \\ &= (\psi(\pi_1(i), Y) - \psi(\pi_2(i), Y))\cdot(g(y_i, Y) - g(y_j, Y)). \end{aligned} \tag{8}$$

This represents the cost incurred by swapping the adjacent instances in π_1. Due to difficulties in obtaining positional information in Procedure 1, we therefore define the importance weight function according to relevance gains.

$$\tilde{w}_{ij}(X, Y) = \left|g(y_i, Y) - g(y_j, Y)\right|. \tag{9}$$

Intuitively, the larger the difference between the relevance gains associated with two different examples, the more important it is to predict the preference between them correctly. Moreover, this choice of weights enjoys sound regret properties which will be proved theoretically in the next section.

The test procedure (Rank_Predict) assigns a preference degree to each instance x_i according to the degree function $f(x_i, X)$, which increases by 1 if x_i is strictly preferred to x_j such that $c(x_i, x_j, X) - c(x_j, x_i, X) = 1$, and $\frac{1}{2}$ if x_i is regarded as equally good as x_j such that $c(x_i, x_j, X) - c(x_j, x_i, X) = 0$. These instances are then sorted in a non-increasing order of the preference degrees.

4. Regret analysis

We now apply the well-established regret transform principle [6,19] to analyze the reduction from subset ranking to binary classification. We first prove a guarantee on the consistency of a reduction when zero regret is attained, and then we provide novel regret bounds when non-zero regret is considered.

4.1. Consistency of reduction methods

We shall rewrite (7) by replacing the original distribution $\mathcal{P}$ with the induced distribution $\tilde{\mathcal{D}}$ due to the reduction:

$$\begin{aligned}\mathcal{L}_{\text{class}}(c, \tilde{\mathcal{D}}) &= E_{\tilde{t}_{ij}\sim\tilde{\mathcal{D}}} l_{\text{class}}(c, \tilde{t}_{ij}) = \frac{1}{Z} E_{S\sim\mathcal{D}} \sum_{(i,j)} l_{\text{class}}(c, \tilde{t}_{ij}) \\ &= E_X L_{\text{class}}(c, X), \end{aligned} \tag{10}$$

where $Z = \frac{n(n-1)}{2}$ is the normalization constant, and

$$\begin{aligned}L_{\text{class}}(c, X) &= \frac{1}{Z} E_{Y|X} \sum_{(i,j)} l_{\text{class}}(c, \tilde{t}_{ij}) \\ &= \frac{1}{Z} \sum_{i,j} E_{(y_i,y_j,Y)|(x_i,x_j,X)} (l_{\text{class}}(c, \tilde{t}_{ij}) + l_{\text{class}}(c, \tilde{t}_{ji})). \end{aligned} \tag{11}$$

Lemma 1. *Given a set $X \in \mathcal{U}$, define the optimal subset preference function $\hat{c} \in \tilde{\mathcal{C}}$ as a minimizer of (11). Let the importance weights be defined as in (9). Then for $\hat{c}(x_i, x_j, X) - \hat{c}(x_j, x_i, X) = 1$, it holds that*

$$E_{(y_i,y_j,Y)|(x_i,x_j,X)} (g(y_i, Y) - g(y_j, Y)) \geq 0.$$

Proof. Note that (11) takes its minimum when each conditional expectation term in the summation achieves its minimum. Substituting (6) and (9) into (11), we have

$$\begin{aligned}&\frac{1}{2} \cdot E_{(y_i,y_j,Y)|(x_i,x_j,X)} [\tilde{w}_{ij}(X, Y) \cdot I(y_i > y_j) \cdot (1 - \hat{c}(x_i, x_j, X) + \hat{c}(x_j, x_i, X)) \\ &\quad + \tilde{w}_{ij}(X, Y) \cdot I(y_j > y_i) \cdot (1 - \hat{c}(x_j, x_i, X) + \hat{c}(x_i, x_j, X))] \\ &= \frac{1}{2} \cdot E_{(y_i,y_j,Y)|(x_i,x_j,X)} [(\hat{c}(x_j, x_i, X) - \hat{c}(x_i, x_j, X)) \cdot (I(y_i > y_j) + I(y_j > y_i)) \\ &\quad \cdot (g(y_i, Y) - g(y_j, Y)) + (I(y_i > y_j) + I(y_j > y_i)) \cdot \tilde{w}_{ij}(X, Y)] \\ &= \frac{1}{2} \cdot [(\hat{c}(x_j, x_i, X) - \hat{c}(x_i, x_j, X)) \cdot E_{(y_i,y_j,Y)|(x_i,x_j,X)} (g(y_i, Y) - g(y_j, Y)) + E_{(y_i,y_j,Y)|(x_i,x_j,X)} \tilde{w}_{ij}(X, Y)]. \end{aligned}$$

Assume by contradiction that $E_{(y_i,y_j,Y)|(x_i,x_j,X)} (g(y_i, Y) - g(y_j, Y)) < 0$. Consider any $k, k' \in \{1, \ldots, n\}$, there exists a preference function $c \in \tilde{\mathcal{C}}$ such that $c(x_k, x_{k'}, X) - c(x_{k'}, x_k, X) = \hat{c}(x_k, x_{k'}, X) - \hat{c}(x_{k'}, x_k, X)$, when $k, k' \neq i, j$ and $c(x_i, x_j, X) - c(x_j, x_i, X) = -1$. Then we get $L_{\text{class}}(c, X) < L_{\text{class}}(\hat{c}, X)$ which stands in contradiction to the subset preference optimality of $\hat{c}$. □

The above lemma together with the result obtained in Theorem 1 allows us to derive the following statement.

Theorem 2. *For one of the measures* AP, NDCG, ERU *and* RBP, *consider position-sensitive subset ranking on X using importance weighted classification. Let the importance weights be defined as in (9). Let* Rank_Predict($\hat{c}$) *be an ordering induced by the optimal subset preference function $\hat{c}$ with respect to $X \in \mathcal{U}$. Then it holds that*

$$L_{\text{rank}}(\text{Rank\_Predict}(\hat{c}), X) = L_{\text{rank}}(\hat{h}, X)$$

where $\hat{h}$ is the optimal subset ranking function with respect to X, i.e. the minimization of the conditional expectation in (3).

The theorem states conditions that lead to a consistent reduction method, in the sense that given an optimal (zero-regret) binary classifier, the reduction can yield a ranker with the minimal expected loss conditioned on X. It is reasonable to extend this result to the optimal subset ranking scenario (see Proposition 1).

4.2. Regret bounds

Here *regret* quantifies the difference between the achieved loss and the optimal loss in expectation. More precisely, let $h^* = \arg\min_h \mathcal{L}_{\text{rank}}(h, \mathcal{D})$ denote the optimal ranking function within $\mathcal{H}$. The regret of ranker h with respect to the distribution $\mathcal{D}$ is defined to be

$$\mathcal{R}_{\text{rank}}(h, \mathcal{D}) = \mathcal{L}_{\text{rank}}(h, \mathcal{D}) - \mathcal{L}_{\text{rank}}(h^*, \mathcal{D}), \tag{12}$$

and the regret of h on the subset X is

$$R_{\text{rank}}(h, X) = L_{\text{rank}}(h, X) - L_{\text{rank}}(\hat{h}, X), \tag{13}$$

where $\hat{h}$ is the optimal subset ranking function as defined previously.

Similarly, let $c^* = \arg\min_c \mathcal{L}_{\text{class}}(c, \tilde{\mathcal{D}})$ denote the optimal preference function within $\mathcal{C}$. The regret of classifier c with respect to the induced distribution $\tilde{\mathcal{D}}$ is defined to be

$$\mathcal{R}_{\text{class}}(c, \tilde{\mathcal{D}}) = \mathcal{L}_{\text{class}}(c, \tilde{\mathcal{D}}) - \mathcal{L}_{\text{class}}(c^*, \tilde{\mathcal{D}}), \tag{14}$$

and the regret of c on the subset X is

$$R_{\text{class}}(c, X) = L_{\text{class}}(c, X) - L_{\text{class}}(\hat{c}, X), \tag{15}$$

where $\hat{c}$ is the optimal subset preference function as defined previously.

Note that $R_{\text{class}}(c, X)$ is scaled by a normalization constant which relies on the total number of induced pairwise preferences, while this is not used in $R_{\text{rank}}(h, X)$. For fairness and simplicity, we leave out the normalization constant Z as defined in Eq. (10), and let $\tilde{R}_{\text{class}}(c, X) = Z \cdot R_{\text{class}}(c, X)$. We then provide an upper-bound that relates the subset ranking regret $R_{\text{rank}}(h, X)$ to the cumulative classification regret $\tilde{R}_{\text{class}}(c, X)$, which can be naturally extended to the ranking regret without set-dependency due to Proposition 1. Before continuing, we need to present some auxiliary results for proving the regret bounds.

Definition 5 (*Proper Pairwise Regret*)**.** Given a set X, for any two instances $x_i, x_j \in X$, we denote the pairwise loss of ordering x_i before x_j by

$$L_{\text{pair}}(x_i, x_j, X) = E_{(y_i, y_j, Y)|(x_i, x_j, X)} \tilde{w}_{ij}(X, Y) \cdot I(y_j > y_i),$$

and denote the associated pairwise regret by

$$R_{\text{pair}}(x_i, x_j, X) = \max(0, L_{\text{pair}}(x_i, x_j, X) - L_{\text{pair}}(x_j, x_i, X)).$$

If $L_{\text{pair}}(x_i, x_j, X) - L_{\text{pair}}(x_j, x_i, X) \geq 0$, then $R_{\text{pair}}(x_i, x_j, X)$ is called proper.

The above definition is parallel to the *proper pairwise regret* defined in [3] with respect to the AUC loss function.

Lemma 2. *Let the importance weights be defined as in (9). For any $i, j, k \in [n]$, if $R_{\text{pair}}(x_i, x_j, X)$ and $R_{\text{pair}}(x_j, x_k, X)$ are proper, then*

$$R_{\text{pair}}(x_i, x_k, X) = R_{\text{pair}}(x_i, x_j, X) + R_{\text{pair}}(x_j, x_k, X).$$

Proof. Since $R_{\text{pair}}(x_i, x_j, X)$ is proper, we have

$$\begin{aligned} R_{\text{pair}}(x_i, x_j, X) &= E_{(y_i, y_j, Y)|(x_i, x_j, X)} \tilde{w}_{ij}(X, Y) \cdot (I(y_j > y_i) - I(y_i > y_j)) \\ &= E_{(y_i, y_j, Y)|(x_i, x_j, X)} (I(y_j > y_i) + I(y_i > y_j))(g(y_j, Y) - g(y_i, Y)) \\ &= E_{(y_j, Y)|(x_j, X)} g(y_j, Y) - E_{(y_i, Y)|(x_i, X)} g(y_i, Y). \end{aligned} \tag{16}$$

Similarly,

$$R_{\text{pair}}(x_j, x_k, X) = E_{(y_k, Y)|(x_k, X)} g(y_k, Y) - E_{(y_j, Y)|(x_j, X)} g(y_j, Y).$$

By combining the above with (16), we obtain the desired result. □

We further introduce Lemma 3 used in our analysis, which was proved in [3].

Lemma 3 (*[3]*)**.** *For any sequence $(a_1, \ldots, a_n)$, let $(a_{(1)}, \ldots, a_{(n)})$ be a sequence sorting the values of $(a_1, \ldots, a_n)$ in a non-increasing order. $\forall i \in \mathbb{N}$, let $\binom{i}{2} = \frac{i \cdot (i-1)}{2}$. If $(a_{(1)}, \ldots, a_{(n)})$ is majorized by $(n-1, \ldots, 0)$, then for any $j \in [n-1]$, it holds that*

$$\sum_{u=1}^{j} \sum_{v=j+1}^{n} I(a_v \geq a_u) \leq 2 \cdot \left(\sum_{v=j+1}^{n} a_v - \binom{n-j}{2} \right).$$

Majorization is originally introduced in [22]. A sequence $(a_1, \ldots, a_n)$ majorizes a sequence $(b_1, \ldots, b_n)$ if and only if $a_1 \geq \cdots \geq a_n$, $b_1 \geq \cdots \geq b_n$ and $\sum_{j=1}^{k} a_j \geq \sum_{j=1}^{k} b_j$ when $k < n$ and $\sum_{j=1}^{n} a_j = \sum_{j=1}^{n} b_j$.

In what follows, we re-index the instances in X according to $\hat{\pi}$, i.e., $j = (\hat{\pi})^{(-1)}(j)$. Taking $\hat{\pi}$ as the target permutation, any permutation π on the same set can be transformed into $\hat{\pi}$ via successive rank-adjacent transpositions [18]. By flipping one discordant pair with adjacent ranks, we get an intermediate permutation. Let $\pi^{(i)}$ denote the intermediate permutation after i transposition operations. For convenience of modeling, we map each discordant pair in the set $\Gamma = \{(v, u) : u < v, \pi(v) < \pi(u)\}$ to the number of adjacent transpositions required to flip it. Specifically, we adopt the transposition strategy of choosing the instance x_j in an increasing order of j and transposing the discordant pairs associated with x_j. More precisely, let $u^- \in \{1, \ldots, u-1\}$ and $u^+ \in \{u+1, \ldots, n\}$, we have

$$i = \sum_{u^-} \tau_1(u^-, \pi) + \sum_{u^+} I(\pi(u^+) < \pi(u)) \cdot I(\pi(v) \leq \pi(u^+)),$$

where $\tau_1(u^-, \pi) = \sum_j I(\pi(j) < \pi(u^-)) \cdot I(u^- < j)$ can be interpreted as the total number of discordant pairs associated with x_{u^-}. In most cases, π is only required to match $\hat{\pi}$ in the top $k (k < n)$ ranks, and the instances ranked below k are regarded as a tie. We adapt the above transposition strategy and the resulting strategy is formulated as follows including two cases.

For the first case, let $\Gamma_1 = \{(v, u) : u < v, \pi(v) < \pi(u); u \leq k, \pi(u) \leq k\}$ denote the set of discordant pairs appearing before k. Given $(v, u) \in \Gamma_1$, we have

$$i = \sum_{u^-} I(\pi(u^-) \leq k) \cdot \tau_1(u^-, \pi) + \sum_{u^+} I(\pi(u^+) < \pi(u)) \cdot I(\pi(v) \leq \pi(u^+)).$$

For the second case, let $\Gamma_2 = \{(v, u) : u < v, \pi(v) < \pi(u); u \leq k, \pi(u) > k; \pi(v) \leq k\}$ denote the set of discordant pairs with only one instance appearing before k. We further define $\Gamma_2^{(1)} = \{(x_v, x_u) \in \Gamma_2 : v \leq k\}$ and $\Gamma_2^{(2)} = \{(x_v, x_u) \in \Gamma_2 : v > k; \tau_2(u, \pi) = \tau_3(v, \pi)\}$, where $\tau_2(u, \pi) = \sum_{u^-} I(\pi(u^-) > k)$ can be interpreted as the total number of instances with smaller indices than x_u and appearing below k, and $\tau_3(v, \pi) = \sum_{k^+} I(\pi(k^+) < \pi(v))$ can be interpreted as the total number of discordant pairs between x_v and those with larger indices than x_k. Given $(v, u) \in \Gamma_2^{(1)} \cup \Gamma_2^{(2)}$, we have

$$i = \sum_{u^-} I(\pi(u^-) > k) \cdot \tau_4(u^-, \pi) + \sum_{u^+} I(\pi(u^+) \leq k) \cdot I(u^+ \leq k) \cdot I(v \leq u^+) + 1 + |\Gamma_1|,$$

where $\tau_4(u^-, \pi) = \sum_j I(\pi(j) < \pi(u^-)) \cdot I(\pi(j) \leq k) \cdot I(u^- < j) - \tau_2(u, \pi)$ can be interpreted as the total number of discordant pairs to be flipped between the instance x_{u^-} and those appearing before k. Equipped with these preparations, we are in a position to prove the upper regret bound for the subset ranking problem.

Theorem 3. *Given one of the measures chosen from* AP, NDCG, ERU *and* RBP, *now consider position-sensitive subset ranking on X using the importance weighted classification. The discount factor d_i for position i is defined as in Section 2.2 and the importance weights as in Eq. (9). Then for any binary classifier c, the following bound holds:*

$$R_{\text{rank}}(\text{Rank\_Predict}(c), X) \leq 2(d_1 - d_2) \cdot \tilde{R}_{\text{class}}(c, X). \tag{17}$$

Proof. Fix c. Let $\text{Rank\_Predict}(c) = h$ and $\text{Rank\_Predict}(\hat{c}) = \hat{h}$. By the definition of $R_{\text{rank}}(h, X)$, we can rewrite the left-hand side of Eq. (17) as

$$R_{\text{rank}}(h, X) = E_{Y|X}(M(\hat{h}(X), Y)) - M(h(X), Y)).$$

We start with the case of NDCG, ERU and RBP measures, and obtain that

$$\begin{aligned}
R_{\text{rank}}(h, X) &= E_{Y|X} \sum_{j=1}^{n} d_j \cdot (g(\hat{\mathbf{y}}_j, Y) - g(\mathbf{y}_j, Y)) \\
&= E_{Y|X} \sum_{(v,u)\in\Gamma} (d_{\pi^{(i)}(u)} - d_{\pi^{(i)}(v)}) \cdot (g(y_u, Y) - g(y_v, Y)) \\
&\leq \max_i (d_{\pi^{(i)}(u)} - d_{\pi^{(i)}(v)}) \sum_{(v,u)\in\Gamma} R_{\text{pair}}(x_v, x_u, X) \\
&\leq (d_1 - d_2) \cdot \sum_{(v,u)\in\Gamma} \sum_{j=u}^{v-1} R_{\text{pair}}(x_{j+1}, x_j, X) \\
&= (d_1 - d_2) \cdot \sum_{j=1}^{n-1} |\{u \leq j < v : \pi(v) < \pi(u)\}| \cdot R_{\text{pair}}(x_{j+1}, x_j, X) \\
&= (d_1 - d_2) \cdot \sum_{j=1}^{n-1} \left[\sum_{u=1}^{j} \sum_{v=j+1}^{n} I(f(x_v, X) \geq f(x_u, X)) \right] \cdot R_{\text{pair}}(x_{j+1}, x_j, X).
\end{aligned} \tag{18}$$

The second equality is due to the fact that

$$M(\hat{\pi}, Y) - M(\pi, Y) = \sum_{i=1}^{\gamma} M(\pi^{(i)}, Y)) - M(\pi^{(i-1)}, Y), \tag{19}$$

where $\pi^{(0)} = \pi$ and $\gamma = |\Gamma|$ denotes the total number of inversions in π (note that $\pi^{(\gamma)}$ is equivalent to $\hat{\pi}$). The second inequality follows by using the fact that the function $(d_j - d_{j+1})$ is monotonically decreasing with j and applying Lemma 2 repeatedly. The third equality follows from algebra, and the fourth from the fact that Rank_Predict outputs a permutation in a non-increasing order of the degree function f.

Next, we discuss the case of AP measure, and obtain that

$$\begin{aligned}
R_{\text{rank}}(h, X) &= E_{Y|X} \sum_{j=1}^{n} \frac{d_j}{\text{rel@}n(\bar{\pi}, Y)} \cdot \left[g(\hat{\mathbf{y}}_j, Y) \cdot \text{rel@}j(\hat{\pi}, Y) - g(\mathbf{y}_j, Y) \cdot \text{rel@}j(\pi, Y) \right] \\
&= E_{Y|X} \sum_{(v,u)\in\Gamma} \frac{d_{\pi^{(i)}(u)} - d_{\pi^{(i)}(v)}}{\text{rel@}n(\bar{\pi}, Y)} \cdot \left[\text{rel@}(\pi^{(i)}(u) - 1)(\pi^{(i)}, Y) + 1 \right] \cdot \left[g(y_u, Y) - g(y_v, Y) \right] \\
&\leq E_{Y|X} \sum_{(v,u)\in\Gamma} (d_{\pi^{(i)}(u)} - d_{\pi^{(i)}(v)}) \cdot (g(y_u, Y) - g(y_v, Y)).
\end{aligned}$$

The second equality is due to (5) and (19). The rest of the derivations follow analogously to the previous cases.

The term on the right-hand side of Eq. (17) can be written as

$$\begin{aligned}
\tilde{R}_{\text{class}}(c, X) &= \sum_{u,v} E_{(y_u, y_v, Y)|(x_u, x_v, X)} (l_{\text{class}}(c, \tilde{t}_{uv}) + l_{\text{class}}(c, t_{vu}) - l_{\text{class}}(\hat{c}, \tilde{t}_{uv}) - l_{\text{class}}(\hat{c}, \tilde{t}_{vu})) \\
&= \frac{1}{2} \cdot \sum_{u,v} E_{(y_u, y_v, Y)|(x_u, x_v, X)} (I(y_u > y_v) + I(y_v > y_u)) \cdot (g(y_u, Y) - g(y_v, Y)) \\
&\quad \cdot \left[(-c(x_u, x_v, X) + c(x_v, x_u, X)) + (\hat{c}(x_u, x_v, X) - \hat{c}(x_v, x_u, X)) \right] \\
&= \frac{1}{2} \cdot \sum_{u,v} \left[(-c(x_u, x_v, X) + c(x_v, x_u, X)) + (\hat{c}(x_u, x_v, X) - \hat{c}(x_v, x_u, X)) \right] \\
&\quad \cdot (E_{(y_u, Y)|(x_u, X)} g(y_u, Y) - E_{(y_v, Y)|(x_v, X)} g(y_v, Y)) \\
&= \frac{1}{2} \cdot \sum_{u<v} (-c(x_u, x_v, X) + c(x_v, x_u, X) + 1) \cdot R_{\text{pair}}(x_v, x_u, X) \\
&= \frac{1}{2} \cdot \sum_{u<v} (-c(x_u, x_v, X) + c(x_v, x_u, X) + 1) \cdot \sum_{j=u}^{v-1} R_{\text{pair}}(x_{j+1}, x_j, X) \\
&= \frac{1}{2} \cdot \sum_{j=1}^{n-1} (2 \cdot |\{u \leq j < v : c(x_v, x_u, X) = 1, c(x_u, x_v, X) = 0\}| \\
&\quad + |\{u \leq j < v : c(x_v, x_u, X) = c(x_u, x_v, X)\}|) \cdot R_{\text{pair}}(x_{j+1}, x_j, X) \\
&= \frac{1}{2} \cdot \sum_{j=1}^{n-1} \left[\sum_{u=1}^{j} \sum_{v=j+1}^{n} \frac{c(x_v, x_u, X) - c(x_u, x_v, X) + 1}{2} \right] \cdot R_{\text{pair}}(x_{j+1}, x_j, X) \\
&= \frac{1}{2} \cdot \sum_{j=1}^{n-1} \left[\sum_{v=j+1}^{n} \sum_{u\neq v} \frac{c(x_v, x_u, X) - c(x_u, x_v, X) + 1}{2} - \binom{n-j}{2} \right] \cdot R_{\text{pair}}(x_{j+1}, x_j, X) \\
&= \frac{1}{2} \cdot \sum_{j=1}^{n-1} \left[\sum_{v=j+1}^{n} f(x_v, X) - \binom{n-j}{2} \right] \cdot R_{\text{pair}}(x_{j+1}, x_j, X).
\end{aligned} \tag{20}$$

The fourth equality follows from Theorem 2 and corresponding properties of algebra. The last equality uses the definition of the degree function f.

Comparing expressions (18) and (20), we obtain the desired bound according to Lemma 3. □

The above theorem derives an upper bound which is up to a constant factor of less than 2 (due to $d_1 - d_2 < 1$) on the regret ratio. This bound is governed by the classification regret on the induced $O(n^2)$ importance weighted binary examples. The following theorem develops a more refined upper bound for subset ranking when the quality is only emphasized on the top $k(k < n)$ rank positions.

Theorem 4. *Given one of the measures chosen from* AP, NDCG, ERU *and* RBP, *now consider position-sensitive subset ranking on X using the importance weighted classification. The discount factor d_i for position i is defined as in Section 2.2 and the importance weights as in Eq. (9). Let α be a constant such that $\alpha < \min_{j' \le k} g(y_{j'}, Y)$, and let $g(y_j, Y) = \alpha$ and $d_j = 0$ for all $j > k$. Then for any binary classifier c, the following bound holds,*

$$R_{\text{rank}}(\text{Rank\_Predict}(c), X) \le 2d_1 \cdot \tilde{R}_{\text{class}}(c, X).$$

Proof. Here we consider the position-sensitive ranking measures together and follow the above mentioned adapted transposition strategy for top-k ranking. Let $\Gamma' = \Gamma_1 \cup \Gamma_2^{(1)} \cup \Gamma_2^{(2)}$ and $\Gamma'' = \{(v, u) : u < v, \pi(v) < \pi(u); u \le k\}$. We obtain

$$\begin{aligned}
R_{\text{rank}}(h, X) &\le E_{Y|X} \sum_{(v,u)\in\Gamma'} (d_{\pi^{(i)}(u)} - d_{\pi^{(i)}(v)}) \cdot (g(y_u, Y) - g(y_v, Y)) \\
&\le \max((d_1 - d_2), d_k) \cdot \left(\sum_{(v,u)\in\Gamma_1} R_{\text{pair}}(x_v, x_u, X) + \sum_{(v,u)\in\Gamma_2^{(1)}} R_{\text{pair}}(x_v, x_u, X) \right) + d_1 \cdot \sum_{(v,u)\in\Gamma_2^{(2)}} R_{\text{pair}}(x_v, x_u, X) \\
&\le d_1 \cdot \sum_{(v,u)\in\Gamma''} \sum_{j=u}^{v-1} R_{\text{pair}}(x_{j+1}, x_j, X) \\
&= d_1 \cdot \sum_{j=1}^{k} |\{u \le j < v : \pi(v) < \pi(u)\}| \cdot R_{\text{pair}}(x_{j+1}, x_j, X).
\end{aligned}$$

The third inequality is due to the facts of $d_1 \ge \max((d_1 - d_2), d_k)$ and $\Gamma' \subseteq \Gamma''$. The rest of the proof follows analogously to Theorem 3. □

The above bound provides guidance for the minimization of top k ranking regret using a binary classifier. This can be implemented by penalizing importance weighted discordant pairs with at least one top k instance, and therefore reducing the number of binary examples from $O(n^2)$ to $O(kn)$.

4.3. Lower bounds

Theorem 3 derives an upper bound which is up to a constant factor of less than 2 (due to $d_1 - d_2 < 1$) on the regret ratio, which extends and improves the previous works in the literature [1,3,12,24]. We will now show the bound is also the best possible by illustrating that the equality given in (17) holds.

Consider a 3-element lower bound example: let the distribution have all its mass on a single 3-element subset $S = \{(x_1, 1), (x_2, 0), (x_3, 0)\}$. We have a classifier c such that $c(x_1, x_2, X) = 0$, $c(x_2, x_1, X) = 0$; $c(x_1, x_3, X) = 1$, $c(x_3, x_1, X) = 0$; $c(x_2, x_3, X) = 1$, $c(x_3, x_2, X) = 0$. Then it is easy to check that for the AP measure $\tilde{R}_{\text{class}}(c, X)$ equals 0.5, and the worst-case value of $R_{\text{rank}}(h, X)$ is 0.5 which is exactly $2(d_1 - d_2) \cdot \tilde{R}_{\text{class}}(c, X)$. In fact, the equality in (17) also holds for the RBP measure in this example.

Consider another 3-element lower bound example: let the distribution have all its mass on a single 3-element subset $S = \{(x_1, 0), (x_2, 1), (x_3, 2)\}$. Suppose a classifier c produces the following estimates: $c(x_1, x_2, X) = 0$, $c(x_2, x_1, X) = 1$; $c(x_1, x_3, X) = 0$, $c(x_3, x_1, X) = 1$; $c(x_2, x_3, X) = 1$, $c(x_3, x_2, X) = 1$. Then it is easy to check that for the NDCG measure $\tilde{R}_{\text{class}}(c, X)$ equals 0.275, and for $R_{\text{rank}}(h, X)$, in the worst case it yields 0.203 which is also exactly $2(d_1 - d_2) \cdot \tilde{R}_{\text{class}}(c, X)$. In fact, the equality in (17) holds for the ERU measure as well in this example.

The above evidences also indicate that Theorem 3 is the best possible.

5. Conclusion

In this paper, we attempt to provide a theoretical analysis supporting subset ranking using binary classifiers. We establish a consistent reduction framework from subset ranking to binary classification, and derive novel tight regret bounds that extend and improve the existing results. Our theoretical analysis reveals the underlying connection between generic subset ranking and binary classification, that is, the improvement of the classification accuracy can reasonably enhance the position-sensitive ranking performance.

Acknowledgments

We wish to thank the anonymous referees for their careful reading and valuable comments. This work was supported in part by the National Natural Science Foundation of China (Grant No. 60621001), the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant Nos. Y1W1031PB1 and GY-110502) and the Project for the National Basic Research 12th Five Program (Grant No. 0101050302).

References

[1] N. Ailon, M. Mohri, An efficient reduction of ranking to classification, in: Proceedings of the 21st Conference on Computational Learning Theory, COLT, 2008, pp. 87–98.
[2] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999.
[3] M.F. Balcan, N. Bansal, A. Beygelzimer, D. Coppersmith, J. Langford, G.B. Sorkin, Robust reductions from ranking to classification, in: Proceedings of the 20th Conference on Computational Learning Theory, COLT, 2007, pp. 604–619.
[4] J. Basilico, T. Hofmann, Unifying collaborative and content-based filtering, in: Proceedings of the 21st International Conference on Machine Learning, ICML, 2004, pp. 65–72.
[5] A. Beygelzimer, V. Dani, T. Hayes, J. Langford, B. Zadrozny, Error limiting reductions between classification tasks, in: Proceedings of the 22nd International Conference on Machine Learning, ICML, 2005, pp. 49–56.
[6] A. Beygelzimer, J. Langford, P. Ravikumar, Error-correcting tournaments, in: Lecture Notes in Computer Science, vol. 5809, 2009, pp. 247–262.
[7] J.S. Breese, D. Heckerman, C. Kardie, Empirical analysis of predictive algorithms for collaborative filtering, in: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, UAI, 1998, pp. 43–52.
[8] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender, Learning to rank using gradient descent, in: Proceedings of the 22nd International Conference on Machine Learning, ICML, 2005, pp. 89–96.
[9] C.J.C. Burges, R. Ragno, Q.V. Le, Learning to rank with non-smooth cost functions, in: Advances in Neural Information Processing Systems, vol. 19, NIPS, MIT Press, 2006, pp. 193–200.
[10] C. Rudin, The P-norm push: a simple convex ranking algorithm that concentrates at the top of the list, Journal of Machine Learning Research 10 (2009) 2233–2271.
[11] C. Cortes, M. Mohri, A. Rastogi, Magnitude-preserving ranking algorithms, in: Proceedings of the 24th International Conference on Machine Learning, ICML, 2007, pp. 169–176.
[12] D. Cossock, T. Zhang, Statistical analysis of bayes optimal subset ranking, in: IEEE Transactions on Information Theory, vol. 54, 2008, pp. 5140–5154.
[13] Y. Freund, R. Iyer, R.E. Schapire, Y. Singer, An efficient boosting algorithm for combining preferences, Journal of Machine Learning Research vol. 4 (2003) 933–969. MIT Press.
[14] R. Herbrich, T. Graepel, K. Obermayer, Support vector learning for ordinal regression, in: Proceedings of the 9th International Conference on Artificial Neural Networks, ICANN, 1999, pp. 97–102.
[15] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems 20 (2002) 422–446.
[16] T. Joachims, Optimizing search engines using clickthrough data, in: Proceedings of the 8th ACM Conference on Knowledge Discovery and Data Mining, KDD, ACM Press, 2002, pp. 133–142.
[17] J.C. Duchi, L.W. Mackey, M.I. Jordan, On the consistency of ranking algorithms, in; Proceedings of the 27th International Conference on Machine Learning, ICML, 2010, pp. 327–334.
[18] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.
[19] J. Langford, A. Beygelzimer, Sensitive error correcting output codes, in: Lecture Notes in Computer Science, vol. 3559, 2005, pp. 158–172.
[20] Q.V. Le, A.J. Smola, Direct optimization of ranking measures, in: CoRR, http://arxiv.org/abs/0704.3359 (2007).
[21] A. Moffat, J. Zobel, Rank-biased precision for measurement of retrieval effectiveness, ACM Transactions on Information Systems 27 (2008) 1–27.
[22] A. Marshall, I. Olkin, Inequalities: theory of majorization and its applications, in: Mathematics in Science and Engineering, vol. 143, 1979, New York.
[23] S. Robertson, H. Zaragoza, On rank-based effectiveness measures and optimization, Information Retrieval 10 (2007) 321–339.
[24] Z.Y. Sun, T. Qin, Q. Tao, J. Wang, Robust sparse rank learning for non-smooth ranking measures, in: 32nd ACM Special Interest Group in Information Retrieval, SIGIR, 2009, pp. 259–266.
[25] T. Pahikkala, E. Tsivtsivadze, A. Airola, J. Järvinen, J. Boberg, An efficient algorithm for learning to rank from preference graphs, Machine Learning 75 (1) (2009) 129–165.
[26] E. Voorhees, Overview of the TREC 2001 question answering track, in: Proceedings of the 10th Text REtrieval Conference, TREC, 2001, pp. 42–51.